Producing ethnicity statistics using GSPREE

Table of Contents

Key Messages of Paper

Purpose

This paper sets out our feasibility research on producing population estimates by ethnic group at local authority level using the Generalised Structure Preserving Estimator (GSPREE) small area estimation technique.

This research is due to be published in January/February 2023 and will feed into the evidence base for the 2023 National Statistician's Recommendation on the future of population and social statistics.

Key Asks of MARP

We are looking for the following from MARP:

- assurance that we have implemented GSPREE appropriately
- confirmation that we have accurately interpreted the results from the GSPREE models and our conclusions are robust
- thoughts on whether GSPREE has potential going forward for producing ethnicity statistics
- ideas for how we can further develop GSPREE to improve the results and overcome the issues that we are seeing
- suggestions for other methods we should be exploring for producing population estimates by ethnic group

Executive Summary

Ethnicity is a high priority topic for users but ONS does not currently produce annual ethnicity statistics at local authority level. This is due to a lack of robust data between censuses. To fill this gap, we are conducting research into the potential to produce statistics on the population by ethnic group using administrative data.

Our initial feasibility research focused on taking a record-level linked data approach. This involved using the Statistical Population Dataset (SPD) V3 as the population base, linking on ethnicity information from a range of administrative datasets and 2011 Census data, using rules to select one ethnicity per person, and of those with a stated ethnicity, producing statistics on the proportion of people in each ethnic group. We refer to these proportions as the admin-based ethnicity statistics. We published our first articles using this approach in August 2021. We published a second iteration in May 2022 following the inclusion of additional data sources and changes to the ethnicity selection rules. Since May, we have been working to incorporate further administrative data sources and intend to publish an update in January/February 2023.

As the research so far has not included adjustments for limitations with the admin data (such as missingness and mis-recording of ethnicity), we are exploring methods in collaboration with ONS methodologists to build on the admin data work. ONS previously published research which investigated using survey data to improve ethnicity estimates from administrative data at a local authority level using the Generalised Structure Preserving Estimator (GSPREE) method. This used English School Census (ESC) data, the 2011 Census and the Annual Population Survey (APS). The research found that the method showed promise for the ability to produce more robust ethnicity estimates to lower geographic levels than could be produced from survey data or ESC data alone. Our work builds on this approach by replacing the single administrative data source (ESC data) with our admin-based ethnicity dataset, which combines data from a number of administrative data sources and has broader coverage.

We have used GSPREE to produce estimates for the 5-category ethnic groups (Asian, Black, Mixed, White, Other) by local authority for 2020. Our inputs into the model were counts of ethnic group by local authority from the 2020 APS data and the 2020 admin-based ethnicity dataset. The 2020 SPD V3 population totals by local authority and weighted national level ethnicity distributions from the 2020 APS were used for benchmarking the GSPREE estimates.

We have compared the GSPREE estimates and the admin-based ethnicity statistics against Census 2021 estimates to assess the approaches. The admin-based ethnicity statistics are closer to Census 2021 than the GSPREE estimates for the majority of local authorities for the Asian, Black, Mixed and White ethnic groups. The GSPREE estimates are closer to the Census 2021 estimates for the majority of local authorities for the Other ethnic group. These results appear to be driven by two key factors:

1) differences in ethnicity estimates between the APS and Census, even at national level. These differences may be driven by the sampling approach and mode of data collection in the APS.
2) the ethnic group selection rules we have implemented for our admin-based ethnicity statistics, which have a large impact on the Other ethnic group.

Given that the GSPREE estimates produced so far do not appear to be an accurate representation of the population (based on comparisons with Census 2021), we are seeking input from the panel on the potential of GSPREE going forward.

Producing ethnicity statistics using GSPREE

Introduction

Ethnicity is a high priority topic for users, particularly for the analysis of inequalities. However, ONS does not currently produce annual estimates of the population by ethnic group at local authority level, with the only official estimates coming from Census data.

As we move towards the 2023 National Statistician’s recommendation on the future of population and social statistics, we are conducting research into the potential to produce estimates of the population by ethnic group from administrative and survey data. The research will provide evidence of our ability to produce ethnicity estimates if Census 2021 is the last of its kind. However, even if the decision is to have another Census in 2031, we are aiming to develop a method to produce annual ethnicity estimates to meet user needs between censuses.

The user needs for ethnicity statistics are as follows:

- annual estimates
- 18 ethnic groups, as per the harmonised standard
- estimates by local authority as a minimum, and ideally for police force areas, health authority boundaries and lower levels of geography
- estimates that are consistent with the mid-year estimates
- the ability to cross-tabulate ethnicity with other variables, such as income and health

We have initially explored using a record-level linked data approach to combine a range of administrative datasets plus 2011 Census data. However, alongside this, we are investigating other methods that could be used to produce accurate ethnicity statistics at low geographic levels. The Generalised Structure Preserving Estimator (GSPREE) is the method we are currently exploring. Alongside this, methodologists within ONS are considering whether there are alternative methods that we should trial in future.

Although there is a user need for the more detailed ethnic group breakdown and for lower levels of geography, we have initially tested GSPREE using the 5-category ethnic group breakdown (Asian, Black, Mixed, White, Other) at local authority level. If successful, we will work with ONS methodologists to expand the approach to a more granular level and explore the ability to produce multivariate statistics.

Background

Between 2001 and 2009, ONS published annual population estimates by ethnic group (PEEGs) for local authorities in England and Wales as experimental statistics. The methodology behind them was to disaggregate by ethnic group the cohort component population accounts for the mid-year estimates for each local authority. The outputs provided breakdowns by quinary age group by sex for England and for

Wales, and by three broad age groups by sex for each local authority. However, concerns were raised about the accuracy of PEEGs following publication of the 2009 estimates and a decision was made in June 2012 to stop further production. The findings from an external review of PEEGs was published in 2017.

In 2017, two research papers were published on using the Generalised Structure Preserving Estimator approach (GSPREE) to produce ethnicity estimates from admin and survey data: one for 2011 and one for 2015. This research used 2011 Census, Annual Population Survey (APS) and English School Census (ESC) data.

In 2017 and again in 2019, research reports were published on producing population estimates by ethnic group using Annual Population Survey (APS) data. In December 2021 experimental statistics that primarily use APS data were published. Due to limitations with the APS data, these were only produced at national and regional level. Alongside this, a review of the current evidence base for population estimates by ethnic group was published.

In August 2021 we published feasibility research on producing statistics on the population by ethnic group from administrative data. These admin-based ethnicity statistics (version 1) were produced by combining English School Census (ESC), Hospital Episode Statistics (HES) and Improving Access to Psychological Therapies (IAPT) data at record-level and using a rules-based approach for dealing with multiple recorded ethnicities for an individual, unknowns and refusals. We then produced figures on the proportion of people in each ethnic group, of those with a stated ethnicity in the admin-based ethnicity dataset. No adjustment was made to account for missingness when calculating the proportions. In May 2022, we published an update to our admin-based ethnicity statistics (version 2), which also incorporated Birth Notifications, Higher Education Statistics Agency (HESA) data and the Emergency Care Dataset (ECDS). Alongside this, we produced a set of figures based on also linking ethnicity information from 2011 Census data into the admin-based ethnicity dataset.

Since then, we have been incorporating additional administrative data sources into our admin-based ethnicity dataset. Individualised Learner Record (ILR) data are being incorporated to improve coverage for England. Welsh School Census (WSC) data, Patient Episode Data for Wales (PEDW) and the Emergency Department Dataset (EDDS) are also being added to expand the coverage to Wales[1]. Alongside this, we have been building on the earlier work of using the GSPREE method to produce ethnicity statistics from combined administrative and survey data.

In January/February 2023, we plan on publishing our updated admin-based ethnicity statistics (version 3) for 2020 for England and separately, for the first time, publishing

[1] Patient Episode Data for Wales (PEDW) and Emergency Department Dataset (EDDS) are no longer included in the admin-based ethnicity statistics or GSPREE estimates planned on being published in January/February 2023 due to quality issues.

our statistics for Wales. Within these papers we also intend to include our GSPREE estimates. These papers will include comparisons with Census 2021 to compare and assess both methods.

Data

The Generalised Structure Preserving Estimator (GSPREE) method combines aggregate data from recent survey estimates with proxy data source(s), usually from administrative or census sources. Row and column margins are used to benchmark the data.

The inputs into our GSPREE method are as follows:

Proxy data - crosstabulation of counts from our admin-based ethnicity dataset for 5-category ethnic group by local authority.

Further information on the data sources and method used to produce our admin-based ethnicity dataset can be found in the Annexe.

Survey data - crosstabulation of unweighted Annual Population Survey (APS) counts for 5-category ethnic group by local authority.

The APS is a continuous household survey, comprising the Labour Force Survey (LFS) supplemented by sample boosts in England, Wales and Scotland to ensure small areas are sufficiently sampled. The APS does not include most people living in communal establishments (such as care homes or prisons) or anyone else living outside private households. Information on some students living in halls of residence is collected where the students' parents live in a sampled household. APS data covering July 2019 to June 2020 were used.

Row benchmarks – Local authority level population totals from the Statistical Population Dataset V3 (SPD)

The Statistical Population Dataset V3 is a record-level dataset produced by linking together a number of administrative datasets and using a set of activity-based rules to determine whether someone is usually resident, meaning they should be retained in the dataset. The numbers of people in the SPD in each local authority were used as the row benchmarks, with no adjustment for under- or over-coverage in the SPD.

Column benchmarks – National weighted ethnicity distributions from the Annual Population Survey (APS) applied to the SPD population total.

The column benchmarks were based on the weighted ethnicity totals from the APS. As the column and row benchmarks need to sum to the same number, the column benchmarks were produced by taking the proportion of people in each ethnic group in the weighted APS data and applying this to the overall SPD population total for England and Wales.

The column benchmark for each ethnic group *e* were therefore calculated as:

$$Benchmark_e = \frac{APS\ tot_e}{APS\ tot} \times SPD\ tot$$

Where:

Benchmark$_e$ is the adjusted column benchmark for ethnic group *e*,

APS tot$_e$ is the weighted estimate of the number of people in ethnic group *e*, in England and Wales,

APS tot$_e$ is the weighted estimate of the number of people in England and Wales,

SPD tot is the estimate of the England and Wales population from the SPD.

Methods

The Generalised Structure Preserving Estimator (GSPREE) is a small area estimation method (Luna-Hernandez et al., 2015) that can be used to produce estimates for a contingency table. Small area estimation methods like GSPREE look to draw strength across multiple data sources, producing more accurate estimates than may otherwise be possible from the sources individually.

A more detailed explanation of the GSPREE methodology is available in Annexe B and at Luna-Hernandez *et al*. (2015). Measures of accuracy are obtained via a bootstrap.

This technique has been used to combine data from multiple sources to produce population estimates by ethnicity at local authority level.

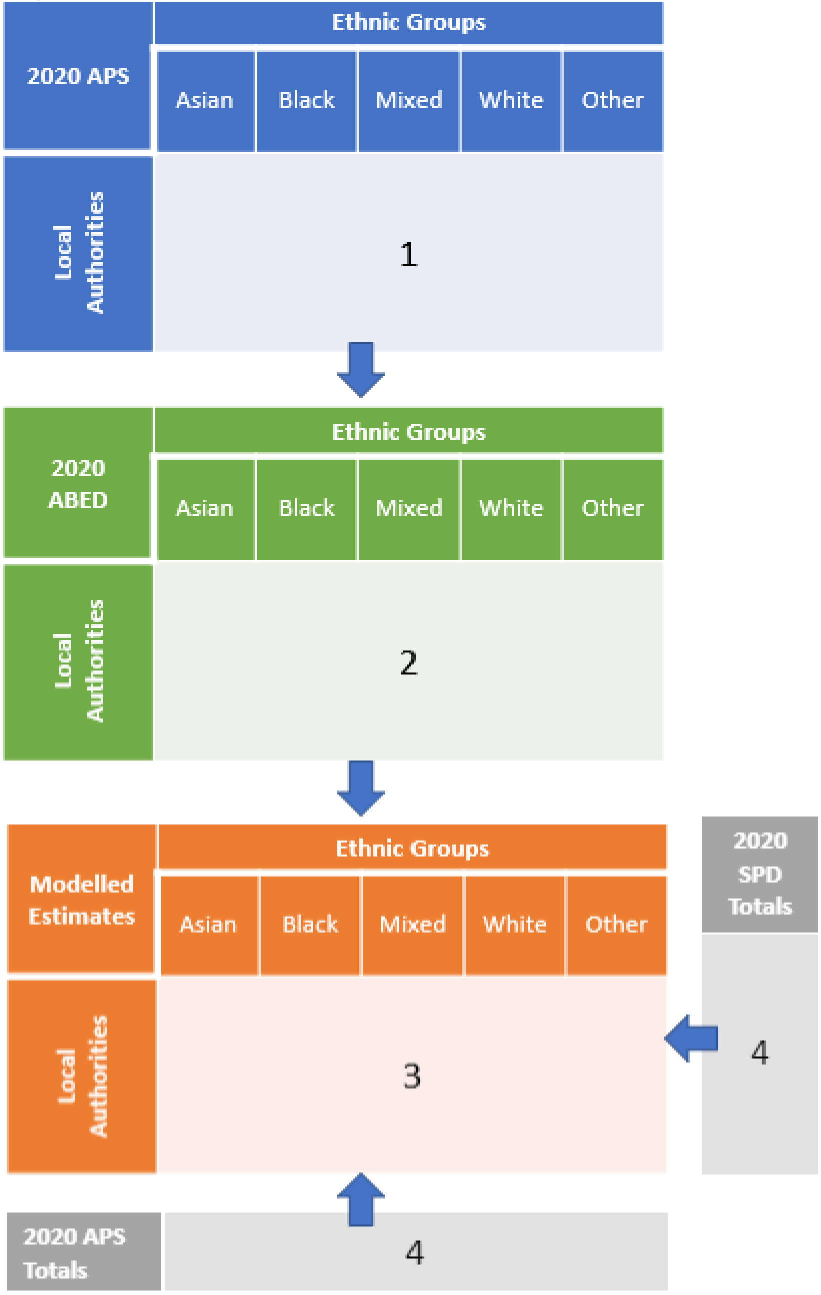
The data structure used by the GSPREE models to update the tables of ethnicity by local authority is shown in Figure 1. A crosstabulation of the unweighted 5-category ethnic group by local authority counts from the Annual Population Survey (APS) is inputted into the model in step one. The GSPREE model used assumes a multinomial distribution for the sampling cell counts in each area, but an alternative approach could be used that allows direct estimates instead. The APS aims to provide up-to-date estimates that are representative of the population, but has too small a sample size to produce results of sufficient quality.

Counts of 5-category ethnic group by local authority produced from the proxy source, the admin-based ethnicity dataset, are also inputted into the model. The admin-based ethnicity data are considered a proxy for the association structure of the population quantities of interest. At step two, the relationship between the association structures of the survey data and the proxy data source is fitted using a log-linear model, and used to produce updated estimates of the target association structure.

From the model relationship between the survey data and the association structure of the proxy, the modelled estimates are compiled in step three. They are compiled by updating the association structure of the proxy through multiplying by an estimated GSPREE model parameter.

Finally at step four, these estimates are adjusted with the row and column benchmarks using iterative proportional fitting to produce final estimates. The row benchmarks are from the 2020 Statistical Population Dataset V3 (SPD). The column benchmarks are Annual Population Survey (APS) 2020 weighted ethnic group distribution applied to the SPD 2020 population.

Figure 1: Current GSPREE data structure for model, 2020



Source: Office for National Statistics

Notes:

1. APS – Annual Population Survey
2. ABED – Admin-based ethnicity dataset
3. SPD – Statistical Population Dataset V3

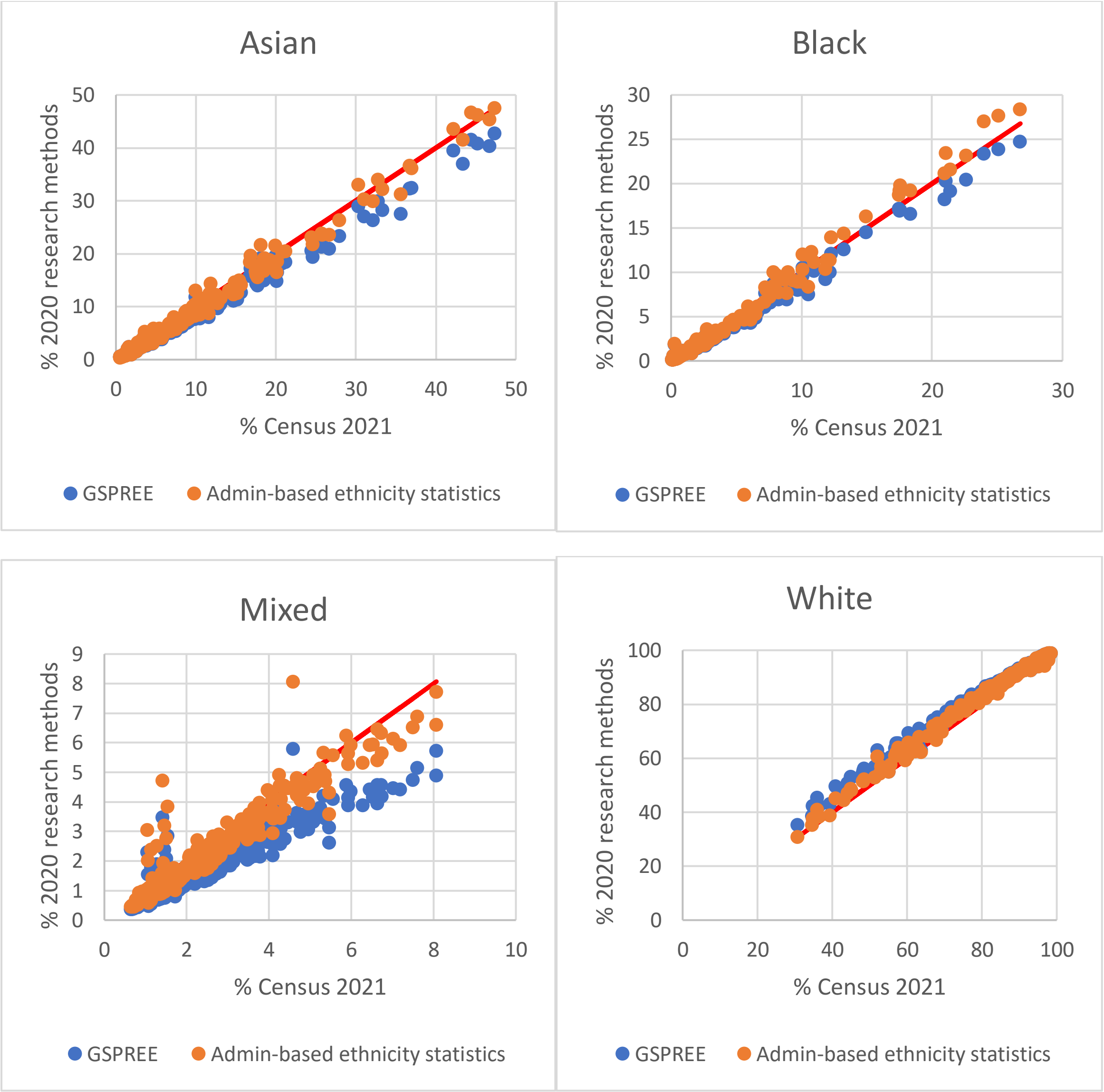
Results

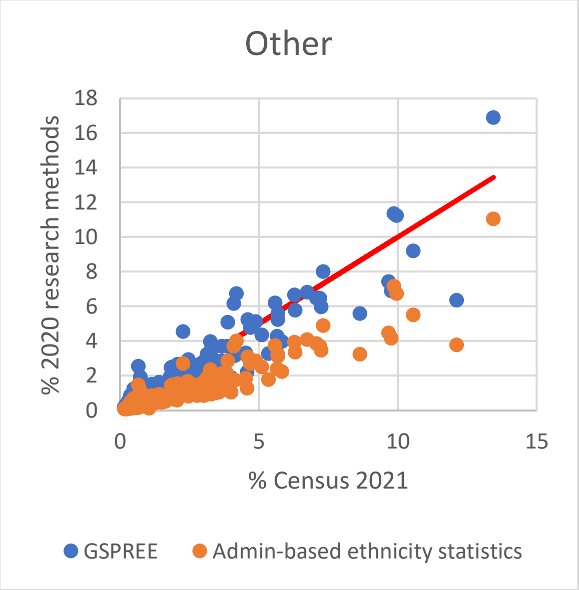
We have produced estimates of the proportion of people in each 5-cateory ethnic group in each local authority using two approaches:

1) the GSPREE method
2) directly taking the proportion of people in each ethnic group out of those with a stated ethnicity in the admin-based ethnicity dataset (referred to as the admin-based ethnicity statistics)

To assess the approaches, we have compared both the GSPREE estimates and the admin-based ethnicity statistics against the Census 2021 estimates (Figure 2).

Figure 2: Proportion of people in each ethnic group in each local authority from the 2020 GSPREE method and the 2020 admin-based ethnicity statistics, versus Census 2021, England and Wales





Source: Office for National Statistics

In addition to plotting the data in the charts above, we have produced two summary measures to assess which set of estimates is closest to the Census for each ethnic group:

1) the proportion of local authorities where the GSPREE estimates are closer to the Census 2021 estimates than the admin-based ethnicity statistics are (Table 1)
2) the sum of the absolute difference in percentage points between the GSPREE/admin-based ethnicity statistics and the Census 2021 estimates (Table 2)

For 91.2% of local authorities, the GSPREE estimates for the Other ethnic group are closer to the Census 2021 proportions than the admin-based ethnicity statistics are. This is likely due to the special rule we implemented for the Any other ethnic group when constructing the admin-based ethnicity dataset (as explained in the data section) to correct for the over-representation of this ethnic group. It appears that these rules have over-corrected the Any other ethnic group in the admin-based ethnicity statistics, bringing the proportions below the Census 2021 proportions for the majority of local authorities.

A similar proportion of local authorities for the Asian and Black ethnic groups are closer to the Census 2021 for the GSPREE estimates and admin-based ethnicity statistics. The sum of the absolute difference in percentage points from each method to Census 2021 shows the admin-based ethnicity statistics to be closer.

The admin-based ethnicity statistics are closer to the Census 2021 estimates for the majority of local authorities for the Mixed and White ethnic groups; this is also reflected in the sum of the absolute difference in percentage points.

Table 1: Proportion of local authorities where the GSPREE estimates are closer to the Census 2021 estimates than the admin-based ethnicity statistics are, England and Wales

Asian	Black	Mixed	White	Other
41.2	48.5	3.9	24.8	91.2

Source: Office for National Statistics

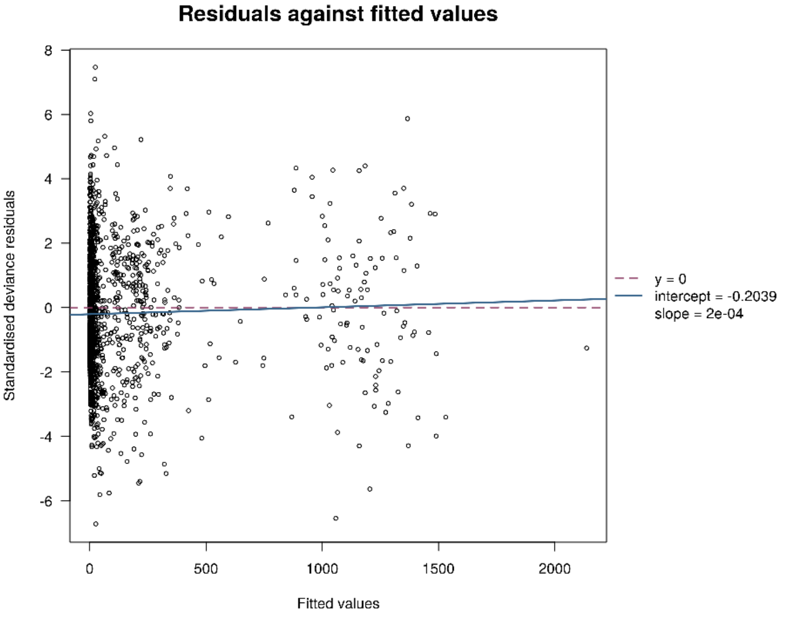
Table 2: The sum of the absolute difference in percentage points between the GSPREE/admin-based ethnicity statistics and the Census 2021 estimates, England and Wales

	GSPREE	Admin-based ethnicity statistics
Asian	397.8	269.9
Black	120.0	108.1
Mixed	283.1	129.3
White	822.4	607.2
Other	118.9	262.8

Source: Office for National Statistics

The figure below presents the model residuals where there is no obvious pattern, suggesting that the model fits well. Influence diagnostics such as Cook's distance and leverage do indicate though that data for the White ethnic group tend to have the highest influence on the model.

Figure 3: GSPREE standardised deviance residuals vs. fitted values



Source: Office for National Statistics

Quality metrics

Measures of accuracy for the GSPREE method can be obtained via resampling methods (e.g. bootstrap), which involve resampling a large number of times from an artificial population.

We have calculated Coefficients of Variation (CVs) and looked at the proportion of local authorities with a CV above 20%. This threshold has been chosen as it has previously been used as an acceptable quality standard for ONS published outputs.

For the Asian ethnic group, the CV is above 20% for 68.2% of local authorities. For the Black, Mixed and Other ethnic groups, the estimates appear to generally be more precise, with approximately 30% of local authorities having a CV above the 20% threshold.

The White ethnic group has no local authorities with a CV above 20%. This is likely due to the high proportion of the population in the White ethnic group.

Table 3: The percentage of local authorities with a coefficient of variation above 20%, England and Wales

Asian	Black	Mixed	White	Other
68.2	29.4	30.0	0.0	32.7

Source: Office for National Statistics

The accuracy measures however do not take into account all quality and coverage issues in the Statistical Population Dataset V3 (SPD). As well as a benchmark for the model, the SPD is also used as the population base for the admin-based ethnicity dataset. The accuracy measures also do not take into account all quality and coverage issues in the admin-based ethnicity dataset which has been used as the proxy source.

Question: Are there alternative statistics that could be used to assess the GSPREE method?

Discussion

APS National

Our GSPREE results show that for the majority of local authorities, the proportion of people in the Asian, Black, Mixed and Other ethnic groups is lower than the Census 2021 proportion. This will be due, in part, to the APS data that is used in the model for benchmarking. The GSPREE method relies on having robust and recent ethnicity data at the national level. There are however limitations with the APS data which impact the estimates that it provides, when compared with Census estimates.

Table 4 shows that the year ending June 2020 APS proportions for the non-white ethnic groups are lower than the Census 2021 estimates. In particular, this is seen for the Asian and Mixed ethnic groups.

Table 4: Ethnic group distributions for England and Wales, year ending June 2020 weighted APS, Census 2021

	Weighted APS 2020	Census 2021
Asian	8.0	9.3
Black	3.7	4.0
Mixed	2.0	2.9
White	84.5	81.7
Other	1.9	2.1

Source: Office for National Statistics

There are several possible factors behind these differences, including:

- APS data are collected across the period of a year, compared with a single reference date for the Census
- the Census relies on people completing a questionnaire themselves whereas an interviewer-led survey is used for the APS
- possible non-response bias by ethnic group in either data source
- the APS is collected at a household level, and a clustering of single ethnicities within households can influence the estimates.
- APS data do not include all communal establishments (some information on students in halls of residence is collected through their families' responses and nursing homes are covered), but they are all represented within Census data
- the pandemic may have had an impact on both the APS and Census 2021 results, however previously comparisons were made between the 2011 Census and the year ending October 2011 APS estimates and similar findings were found

Questions:

Are the APS estimates at a national level robust enough to be used within the GSPREE method?

If not, do the panel have any suggestions for alternative national ethnicity totals that could be used as the column benchmarks?

Inclusion of 2011 Census in the admin-based ethnicity dataset

The admin-based ethnicity dataset has had 2011 Census data linked in as one of the ethnicity data sources. This is because we want to make the best use of all available data sources and the 2011 Census is the most complete source of ethnicity data as at Census Day. By incorporating the 2011 Census data, the proportion of people in the SPD with a stated ethnicity is increased from 80.2% to 84.9% for England and from 41.7% to 71.2% for Wales.

For consistency with our admin-based ethnicity statistics, for our GSPREE model, we have used the version of the admin-based ethnicity dataset which incorporates the 2011 Census data (model 1). A version of the dataset which has been produced only with the administrative data sources is however also available.

We have explored running the GSPREE model with the ethnicity distributions by local authority from the dataset that does not use 2011 Census data as the proxy source (model 2). The results from the two models are broadly similar but we found that for the majority of local authorities, the proportions were closer to the Census 2021 proportions using model 1 compared with model 2. We therefore propose that we continue to incorporate 2011 Census data and would intend on incorporating Census 2021 data as a source in future if the GSPREE method was implemented.

Question: Is it acceptable for Census data to be incorporated as a data source in the admin-based ethnicity dataset that underlies the proxy data input?

How data sources are included in the model

The proxy source that has been used for the model is the admin-based ethnicity dataset which combines multiple administrative data sources and 2011 Census data, as described in the data section. An alternative approach would be to take each source individually and input them into the method as separate proxy sources (Luna-Hernandez et al., 2015). This approach constructs the proxy structure using a convex linear combination of the two available structures. Their association structures are combined using weights (δ and $1 - \delta$, where $0 \leq \delta \leq 1$), which are found via numerical optimisation, as the value that minimizes the deviance of the fitting of the model. We are currently limited by the code we have available which only allows for up to two sources to be used, so we have not been able to explore the impact this would have on the model results.

We have however produced a model which uses the version of the admin-based ethnicity dataset without Census dataset as a proxy source and 2011 Census as a separate proxy source. We found that having 2011 Census as a separate proxy produced similar proportions to when the admin-based ethnicity dataset with 2011 Census incorporated was used as the only proxy source.

Questions:

Is it acceptable to incorporate the different data sources as one combined proxy source (the admin-based ethnicity dataset) or would it be better for them to be inputted as separate proxy sources?

If they were inputted as separate proxy sources, we would still need to consider the following –

- should all data up to the point of interest be used to produce the crosstabulation of ethnicity by local authority or should just a single year of data be used for each source?
- some people have multiple records within a data source so assuming we want just one record per person, we would need an approach for deduplicating the data. Are our current ethnicity selection rules, outlined in Annexe A, a suitable approach or do the panel have other suggestions?
- each data source will include people who are not in the usually resident population. Should we continue to link the data sources to the Statistical Population Dataset and remove those who don't link, before producing the crosstabulation of ethnicity by local authority, or just include anyone present in each admin data source?

Combining England and Wales into one model

We have produced a single model for England and Wales, which means that the GSPREE estimates are benchmarked to the England and Wales ethnicity totals. Consequently, the GSPREE estimates for local authorities in Wales may be influenced by the ethnicity distributions for England. This is however consistent with how we have treated different regions of England, which will have different ethnic compositions; we have not produced separate regional models to account for this.

There are some additional factors to consider for Wales. Most of the datasets that have been used in the production of the admin-based ethnicity dataset either cover England or Wales. This means that the admin-based proxy input for Wales has been largely produced from different datasets to the admin-based proxy input for England. Due to this, and the fact that the Welsh hospital data only go back to 2019 whereas the English hospital data go back to 2009, the proportion of people with a stated ethnicity in the admin-based ethnicity dataset is generally lower in local authorities in Wales compared to local authorities in England. We are therefore considering whether it would be better to run separate models for England and for Wales.

Question: Should we run separate models for England and for Wales?

Adding groupings

The GSPREE model can be made more complex by disaggregating the proxy and survey data into groups. This means the model is fitted independently for each of the groups but is then benchmarked to the national total. If the panel think that GSPREE has potential going forward for producing admin-based ethnicity statistics, we could explore whether this improves the GSPREE estimates. Potential groupings we have considered are regions, age groups, and splitting the local authorities based on the proportion in the white ethnic group or an alternative local authority grouping.

Question: Should we incorporate groupings into the model and if so, are there any suggestions on how we should group the data?

Question: Are there other changes that we can make to the implementation of GSPREE to improve the resulting estimates?

Conclusion

Initial exploration of the proportion of people in each ethnic group in the GSPREE estimates and admin-based ethnicity statistics, compared with the Census 2021 estimates, suggests that the GSPREE approach, while it is relying on Annual Population (APS) data, does not improve our statistics on the population by ethnic group at local authority level. We welcome input from the panel on the potential of GSPREE going forward, particularly on whether it is worth trialling some of the tweaks to the model outlined above, or whether the issues with the APS data are too fundamental.

Future Steps

If the GSPREE method is considered to have the potential to produce estimates of the population by ethnic group at local authority level for 5 ethnic groups, we would plan on looking to expand the approach to a greater number of ethnic groups and also to lower geographic levels. We would also explore the ability to use the method for producing multivariate statistics of ethnicity by other characteristics.

Question: How can we improve the application of GSPREE to enable further breakdowns by ethnic group and/or geography?

To better understand the quality of the admin-based ethnicity dataset that is being used as the proxy data source, we are planning to conduct record-level comparisons between our admin-based ethnicity dataset and the Census 2021 data. Through these comparisons, we will be able to build our understanding of the impact of missingness in the admin data and assess the accuracy of our derived ethnicity variable. This will, in particular, help us to understand the under-representation of the

Other ethnic group and explore whether this can be improved by adapting our ethnicity selection rules.

The Labour Force Survey (which the Annual Population Survey is based on) is currently undergoing transformation. The transformed LFS is due to have a larger sample size than the APS. Once the data are available, we can use transformed LFS data for both the ethnicity by local authority survey input and for the column benchmarks. We will then be able to assess whether this improves the GSPREE outputs or whether the issues outlined above for the Asian, Mixed and White ethnic groups in particular, persist.

Methodology are currently conducting a review of other methods that could be used to produce statistics on population characteristics. We plan to test out the methods that are identified and assess their suitability for producing ethnicity estimates by local authority and lower levels of geography.

Question: Do the panel have suggestions for other methods that we should explore for producing population estimates by ethnic group?

References

Luna-Hernandez, A., Zhang, L., Whitworth, A. and Piller, K. (2015) Small Area Estimates of the Population Distribution by Ethnic Group in England: A Proposal Using Structure Preserving Estimators. Statistics in Transition New Series and Survey Methodology Joint Issue: Small Area Estimation 2014. Vol. 14, No.4, pp. 585-602.

Office for National Statistics (ONS), released 25 August 2017, ONS website, Article, Research Outputs: An approach for estimating ethnicity from survey and administrative data, 2011 - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), revised 25 August 2017, ONS website, Article, Research report on population estimates by characteristics - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), released 20 September 2017, ONS website, Article, Population estimates by ethnic group (PEEGs) – external review - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), released 6 November 2017, ONS website, Article, Research Outputs: Ethnicity estimates from survey and administrative data, 2015 - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), released 4 December 2019, ONS website, Article, Research report on population estimates by ethnic group and religion - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), released 27 July 2020, ONS website, Article, Admin-based population estimates and statistical uncertainty - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), released 27 July 2020, ONS website, Article, Population and migration statistics system transformation – recent updates - Office for National Statistics

Office for National Statistics (ONS), released 6 August 2021, ONS website, Article, Admin-based ethnicity statistics for England, feasibility research - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), released 16 December 2021, ONS website, Article, Population estimates by ethnic group and religion, England and Wales - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), released 16 December 2021, ONS website, Article, Review of the current evidence base for population estimates by ethnic group - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), released 23 May 2022, ONS website, Article, Developing admin-based ethnicity statistics for England - Office for National Statistics (ons.gov.uk)

Office for National Statistics (ONS), released 23 May 2022, ONS website, Article, Producing admin-based ethnicity statistics for England: changes to data and methods - Office for National Statistics (ons.gov.uk)

Annexes

Annexe A – Method for producing the admin-based ethnicity dataset

The 2020 admin-based ethnicity dataset was produced by combining a number of administrative data sources. These were:

- English School Census (ESC), 2011 to 2020
- Hospital Episode Statistics (HES), 2009 to 2020
- Improving Access to Psychological Therapies (IAPT), 2012 to 2018
- Higher Education Statistics Agency (HESA), 2010 to 2020
- Birth Notifications, 2006 to 2020
- Emergency Care Dataset (ECDS), 2020
- Individualised Learner Record (ILR), 2008 to 2020
- Welsh School Census (WSC), 2011 to 2020
- Patient Episode Data for Wales (PEDW), 2019 to 2020[2]
- Emergency Department Dataset (EDDS), 2019 to 2020[2]
- 2011 Census

Ethnicity records from these data sources up to 30 June 2020 were linked to the 2020 Statistical Population Dataset V3 (SPD) based on a unique identifier. Records that did not link to the SPD were dropped. As people may appear multiple times within and across the data sources, sometimes with different ethnicities recorded for them, we implemented a rules-based approach for selecting one ethnicity per person. This was largely based on taking the most recently recorded ethnicity for an individual, with the following special cases, as outlined in our methods publication:

- if someone's most recent ethnicity record says unknown, take their last stated ethnicity if available

- if their most recent ethnicity record says refused, record their final ethnicity as refused.

- if someone's most recent ethnicity record says Any other ethnic group or White not specified (which is a category we created due to some people in HESA just having their ethnicity recorded as White), take a previously recorded ethnicity if available. These were new rules introduced for version 2 of the research. This is because we found that the proportion of people in the Other ethnic group was much higher in the admin-based ethnicity statistics than was expected based on comparisons with 2011 Census. Research within the health sector has suggested that there is over-use of the Any other ethnic group code. Given that our admin-based ethnicity statistics largely rely on health data, this overuse was likely to be contributing to the higher proportion. For White not specified, the additional step was introduced to see whether we

[2] Patient Episode Data for Wales (PEDW) and Emergency Department Dataset (EDDS) are no longer included in the admin-based ethnicity statistics or GSPREE estimates planned on being published in January/February 2023 due to quality issues.

could get a more specific ethnic group for these people, to align to the 18-category ethnic groups that we are aiming to produce ethnicity statistics for.

- if someone has multiple recorded ethnicities on the latest date, record their final ethnicity as unresolved and group them in with those with an unknown ethnicity in the outputs. The exception to this is: if any of the ethnicities are Any other ethnic group and/or White not specified and removing them leaves a single stated ethnicity, take that stated ethnicity instead.

After deriving a final ethnicity for as many people as possible, the 5-category ethnic group variable was derived and cross-tabulated with the local authority variable to produce the proxy data input table. The 2020 admin-based ethnicity dataset has a stated ethnicity recorded for 84.9% of individuals in the SPD for England and 71.2% for Wales.

The remaining individuals in the SPD without a stated ethnicity consists of:

- those in at least one of the administrative data sources but the most recent ethnicity is refused - 10.3% of the SPD for England and 22.9% for Wales
- those in at least one of the administrative data sources but with an unknown ethnic group – 0.8% of the SPD for England and 0.8% for Wales
- those that could not be linked to any of the administrative data sources – 4.1% of the SPD for England and 5.1% for Wales

Annexe B – Further detail on how the GSPREE method works

For a target contingency table *Y* with A rows and J columns, the GSPREE model takes the form:

$$logY_{aj} = \gamma_a + \lambda_j + \beta\alpha_{aj}^X$$

Both the γ_a and λ_j terms, for $a = 1, \dots, A$, $j = 1, \dots, J$, are nuisance parameters based on the proxy table *X*. The term α_{aj}^X represents the association structure of the proxy table.

The GSPREE is characterised by the structural equation:

$$\alpha_{aj}^Y = \beta\alpha_{aj}^X$$

for $a = 1, \dots, A$, $j = 1, \dots, J$.

It therefore proposes to use β to update the association structure of the proxy table.